On the Use of Population Attributable Fraction to Determine Sample Size for Case-Control Studies of Gene-Environment Interaction

Quanhe Yang,[1] Muin J. Khoury,[2] J. M. Friedman,[3] and W. Dana Flanders[4]

Abstract: Most methods for calculating the sample size needed to detect gene-environment interactions use odds ratios to measure the effect size. We show that for any combination of susceptible genotype prevalence and exposure prevalence and their associated risks, the odds ratio measuring strength of interaction corresponds to a population attributable fraction (PAF) because of interaction and vice versa. Simultaneous consideration of odds ratio for interaction and the associated PAF attributable to interaction provides additional insight to investigators evaluating the feasibility and public health relevance of a proposed study.We considered gene-environment interactions on a multiplicative scale, and assumed a dichotomous environmental exposure variable and a single two-allele disease-susceptibility locus. Our results show, for example, that for studies of exposures and genotypes that are common in a population (30%–50%), the PAF for interaction is large (>27%) even if the odds ratio for interaction is only moderate (~2). If simultaneous estimates of interaction odds ratio and PAF indicate that the PAF is so large as to be implausible, the investigator may decide to reevaluate the study design based on detecting a more reasonable PAF. In this case, the associated odds ratio for interaction will be weaker and a considerably larger sample size may be needed.
(EPIDEMIOLOGY 2003;14:161–167)

Key words: gene-environment interaction, sample size, population attributable fraction, case-control study.

Genetic factors contribute to virtually every human disease, conferring susceptibility or resistance, or influencing interaction with environmental factors. The concept of gene-environment interaction is, therefore, a central theme in genetic epidemiologic studies.[1] In recent years, increasing numbers of genetic epidemiologic studies have examined the role of gene-environment interaction in disease etiology.[2–7]

Methods for defining and measuring interactions in epidemiologic studies have been widely discussed.[8–11] From a statistical perspective, interaction is measured as departure from a multiplicative model and is calculated simply as the coefficient of the product of the relative risks of each component factor.[9] From a biological perspective, interaction occurs when two factors both participate in the same mechanism of disease causation and can be measured in terms of departure from an additive model.[12] Although we realize that additive interactions may provide important insights into underlying pathogenic mechanisms, we deal here with the more commonly used multiplicative scale interactions.

When designing a study to detect the effect of gene-environment interactions, investigators need to consider sample size and power. Most methods for calculating sample size use the odds ratio (OR) to measure the strength of gene-environment interactions.[13–16] Other studies have shown the usefulness of population attributable fraction (PAF) as a measure of association in sample size estimation for single exposure variables.[17]

The present study examines the relation of OR for interaction and the associated PAF for interaction as an aid in determining sample size for investigations of gene-environment interactions. We show that, for any combination of susceptible genotype prevalence and exposure prevalence and their associated risks, the OR measuring strength of interaction corresponds to a PAF

Editors' note: *An invited commentary on this article appears on page 137.*

From the [1]National Center on Birth Defects and Developmental Disabilities and [2]Office of Genomics and Disease Prevention, Centers for Disease Control and Prevention, Atlanta, GA; [3]Department of Medical Genetics, University of British Columbia, Vancouver, BC, Canada; and [4]Department of Epidemiology, School of Public Health, Emory University, Atlanta, GA.

Address correspondence to: Quanhe Yang, National Center on Birth Defects and Developmental Disabilities, Centers for Disease Control and Prevention, 4770 Buford Hwy, MS F-45, Atlanta, GA 30341; qyang@cdc.gov

Submitted 28 August 2002; final version accepted 18 September 2002.

because of interaction. Considering the PAF for interaction as well as the OR for interaction in the design phase allows the investigator to reconcile expectations for the effect size of a gene-environment interaction with an assessment of the associated public health impact. We examine how these two measurements are related, and how they can be used to help determine the minimum sample size required to detect a gene-environment interaction in case-control studies.

Methods

Population attributable fraction (also called attributable risk, population attributable risk proportion or etiologic fraction) is defined as the proportion of the disease cases in a population that would be prevented if an exposure were eliminated, assuming the exposure to be causal.[18] For a single binary exposure risk factor, we can define the PAF as:

$$PAF = \frac{P_x(\mathrm{RR} - 1)}{P_x(\mathrm{RR} - 1) + 1} \qquad (1)$$

where P_x is the proportion of exposure in the population and *RR* is the risk ratio associated with that risk factor. Several other formulas can be used to estimate PAF,[18] but this definition, originally proposed by Levin,[19] has been widely used. With an appropriate design, P_x can be estimated among control subjects and *RR* can be replaced by the odds ratio,[20] so all parameters are estimable from a case-control study.

For the study of gene-environment interactions, we assume a dichotomous environmental exposure variable ($e = 1$, exposed, and $e = 0$, absent) and a single dominant disease-susceptibility allele ($g = 1$, present, and $g = 0$, absent). Let R_{ij} be the disease risk among persons with a particular combination of environmental risk factor ($e = 0,1$) and susceptibility genotype ($g = 0,1$), and P_{ij} indicates the proportion of the population with the combination i, j of e and g. We define the population attributable fraction attributable to interaction on a multiplicative scale as:

$$PAF_i = \frac{P_{11}\left(R_{11} - \frac{R_{10}R_{01}}{R_{00}}\right)}{\sum P_{ij}R_{ij}} = \frac{P_{11}(\mathrm{RR}_{11} - \mathrm{RR}_{10}\mathrm{RR}_{01})}{\sum P_{ij}\mathrm{RR}_{ij}} \qquad (2)$$

where R_{ij} and $\mathrm{RR}_{ij} = R_{ij}/R_{00}$ represent the absolute risk and risk ratio for the disease, respectively. P_{11} is the proportion of the population exposed to e and with genotype g simultaneously, $\Sigma P_{ij}R_{ij}$ is the overall risk of the disease in the population, and $R_{10}R_{01}/R_{00}$ would be the risk among those who are exposed to the environmental risk factor and have the susceptible genotype under a multiplicative model. Similar to the interpretation of AF for a single exposure variable, the PAF_i is the proportional excess of disease attributed to the interaction of exposure to environmental risk factor and the susceptible genotype over that which would have occurred if the susceptible genotype and exposure had acted independently, according to a multiplicative model. PAF_i is zero when $\mathrm{RR}_{10}\mathrm{RR}_{01} = \mathrm{RR}_{11}$. If $\mathrm{RR}_{11} < \mathrm{RR}_{10}\mathrm{RR}_{01}$, the value of PAF_i will be negative.

The PAF_i can be estimated by using parameters from a case-control study. In a case-control study of a gene-environment interaction, the effects of the genotype alone, the environmental exposure alone, and the gene-environment interaction can be evaluated in a $2 \times 2 \times 2$ table classified by the presence or absence of the exposure and of the susceptible genotype (see Appendix table).[21] The gene-environment interaction on a multiplicative scale is defined as $\mathrm{RR}_i = \mathrm{RR}_{11}/\mathrm{RR}_{10}\mathrm{RR}_{01}$, *ie*, the factor by which the OR for those exposed to the environmental risk factor and having the disease-susceptibility genotype differs from the product of the effects of the environmental exposure and the susceptible genotype individually. With a case-control study of gene-environment interaction designed so that the odds ratio estimates the corresponding risk ratio,[20] one can estimate PAF_i by substituting this definition of RR_i into Eq 2:

$$PAF_i = \frac{P_{11}R_{10}\mathrm{RR}_{01}(\mathrm{RR}_i - 1)}{P_{11}\mathrm{RR}_i\mathrm{RR}_{10}\mathrm{RR}_{01} + P_{10}\mathrm{RR}_{10} + P_{01}\mathrm{RR}_{01} + P_{00}} \qquad (3)$$

where P_{ij} indicates the proportion of population with the combination i, j of the environmental risk factor and disease-susceptibility genotype, and RR_{ij} is the risk ratio among persons exposed to that combination of environmental risk factor and susceptible genotype.

To estimate minimum sample size required to detect the gene-environment interaction in a case-control study, one needs to specify a set of parameters, *eg*, {P_e, P_g, RR_{10}, RR_{01} and RR_i}, the case-control ratio, and the type I and II errors,[22] where P_e is the population prevalence of exposure to the environmental risk factor and P_g is the population prevalence of the disease-susceptibility gene. Assuming independence of P_e and P_g in the population, we have $P_{11} = P_eP_g$, $P_{10} = P_e(1-P_g)$, $P_{01} = P_g(1-P_e)$ and $P_{00} = (1-P_e)(1-P_g)$.

We used the following formula to estimate the sample size required to detect a gene-environment interaction[23]:

$$N = \frac{(Z_{\alpha/2}\sqrt{\nu_N} + Z_\beta\sqrt{\nu_A})^2}{(\log\mathrm{RR}_i)^2} \qquad (4)$$

where RR_i is a measure of gene-environment interaction effect and $Z_{\alpha/2}$ and Z_β are normal deviates to give a

two-sided significance test at level α with power $1-\beta$. v_N and v_A are proportional to the variance of the logarithm of RR_i under the null hypothesis and under an alternative hypothesis, respectively. A method for calculating v_N and v_A is described in the Appendix.

Straightforward mathematic manipulation of Eq 3 gives

$$RR_i = \frac{RR_{10}RR_{01} + PAF_i(P_{10}RR_{10} + P_{01}RR_{01} + P_{00})}{P_{11}RR_{10}RR_{01}(1 - PAF_i)} \quad (5)$$

For any combination of susceptible genotype, prevalence of exposure and their associated risks (RR_{10} and RR_{01}), a given PAF_i determines RR_i and vice versa. There are infinite ways to specify the combination of parameters needed to estimate sample size, and an investigator simply needs to express an available parameter set in terms of PAF_i to calculate the sample size needed to produce a specified PAF_i. For example, the sample size needed to detect a gene-environment interaction can be calculated from parameter set {P_e, P_g, RR_{10}, RR_{01}, PAF_i} by substituting Eq 5 for RR_i in Eq 4.

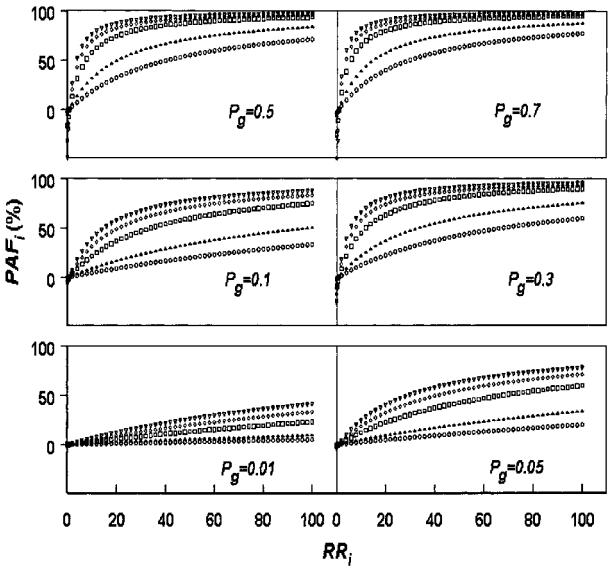

FIGURE 1. Relation of PAF_i to RR_i for various population frequencies of a susceptibility genotype (P_g) and environmental exposure (P_e). The graphs illustrate a type I gene-environment interaction with $RR_{01} = 1.0$, $RR_{10} = 1.0$ and $RR_i > 1.0$. (P_e ○ = 0.05, △ = 0.1, □ = 0.3, ◇ = 0.5, ▽ = 0.7)

Results

PAF_i, the population attributable fraction resulting from a gene-environment interaction, is a function of the population frequencies of the susceptibility genotype (P_g) and the exposure (P_e) as well as of RR_i, the risk ratio for the gene-environment interaction among persons with the susceptible genotype who are also exposed to the environmental risk factor. The type of gene-environment interaction also influences these relations, which are shown in Eqs 3 and 5 above. Three types of gene-environment interaction that cover a wide range of realistic scenarios are:

- type I interactions, where neither the genotype alone nor the exposure alone causes excess risk ($RR_{10} = RR_{01} = 1$) but $RR_{11} > 1$ and $PAF_i > 0$;
- type II interactions, where $RR_{10} > 1$, $RR_{01} = 1$ and $RR_{11} > RR_{10}$ and $PAF_i > 0$; and
- type III interactions, where $RR_{10} > 1$, $RR_{01} > 1$ and $RR_{11} > RR_{10}RR_{01}$ and $PAF_i > 0$.[24]

The sample size estimation remains unchanged within each type of interaction where the effects of RR_{10} and RR_{01} are interchanged, *eg*, sample size requirements for a type II interaction where $RR_{10} = 3$, $RR_{01} = 1$ and $RR_{11} = 5$ equal those where $RR_{10} = 1$, $RR_{01} = 3$ and $RR_{11} = 5$ because of the symmetric effect of RR_{10} and RR_{01} on sample size estimation.

Figure 1 illustrates the relation between PAF_i and RR_i for various values of P_g and P_e for a type I gene-environment interaction. PAF_i increases as RR_i increases, as P_e increases and as P_g increases. The effects are similar with type II or type III gene-environment interactions but are less symmetric with respect to their dependence on P_e and P_g, as expected.

The critical effect on PAF_i of the population frequencies of the susceptibility genotype and the exposure is shown in Figure 2, in which PAF_i is plotted against P_e and P_g when RR_i is held constant ($RR_i = 2.0$, type I interaction). Comparison of Figure 2 with Figure 3, which is an analogous plot of $\log RR_i$ when PAF_i is held constant ($PAF_i = 10\%$, type I interaction), dramatically illustrates the difference between viewing a gene-envi-

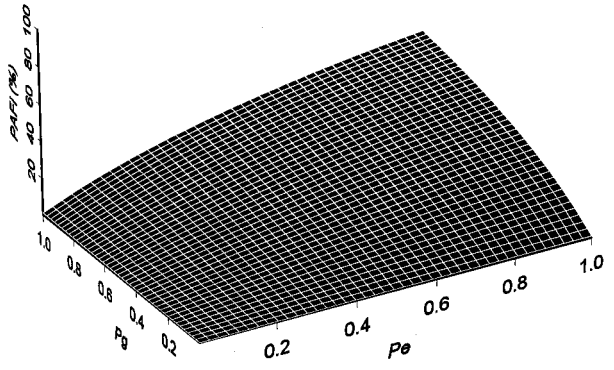

FIGURE 2. Relation of PAF_i to the population frequencies of a susceptibility genotype (P_g) and environmental exposure (P_e). The graph illustrates a type I gene-environment interaction with $RR_{01} = 1.0$, $RR_{10} = 1.0$ and $RR_i = 2.0$.

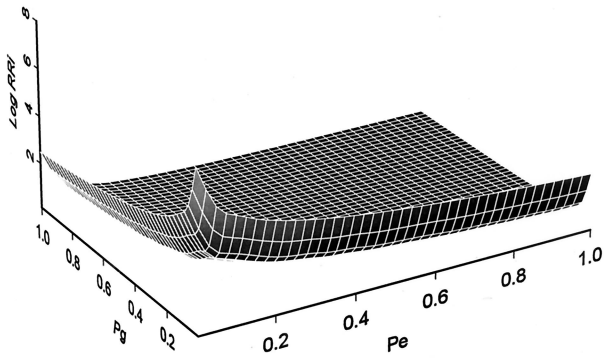

FIGURE 3. Relation of RR_i to the population frequencies of a susceptibility genotype (P_g) and environmental exposure (P_e). Note that RR_i is plotted on a logarithmic scale. The graph illustrates a type I gene-environment interaction with RR_{01} = 1.0, RR_{10} = 1.0 and PAF_i = 10%.

ronment interaction in terms of its effect on RR_i and its effect on PAF_i.

This difference is reflected in the sample size that is required for a case-control study of a gene-environment interaction. The minimum sample size for any given RR_i occurs when the exposure prevalence and the susceptible genotype frequency both lie in the range of about 30% to 50%. This pattern is consistent with the findings of other studies.[13–15,22] When sample size is estimated on the basis of RR_i, the number of cases required becomes smaller, and the PAF_i becomes greater as RR_i increases if other factors remain constant. In contrast, when estimated on the basis of PAF_i, the minimum sample size for any given combination of RR_{10} and RR_{01} occurs when the prevalence of both the exposure and the susceptible genotype are relatively low. If both the exposure prevalence and the susceptible genotype frequency are very low, the sample size required is greater than if both frequencies are less extreme.

The minimal sample sizes for desirable values of PAF_i are often associated with values of RR_i that are unrealistically high. For any given value of PAF_i, increasingly larger values of RR_i are associated with lower frequencies of exposure and/or of the susceptible genotype, other factors being equal (Figure 3). The sample size required increases rapidly as the prevalence of exposure or susceptible genotype frequency becomes more common.

For fixed values of the parameters {P_e, P_g, and PAF_i}, sample size is smaller for type I interactions than for type II or III interactions. This is expected because the associated RR_i decreases as either RR_{10} or RR_{01} increases, other factors being equal. In contrast, when the parameters {P_e, P_g, and RR_i} are fixed, the required sample size is more similar for the three types of interaction.

TABLE 1. Examples of Sample Size Calculations Based on Odds Ratio (RR_i) and Population Attributable Fraction (PAF_i) for Gene-Environment Interaction from Two Recent Case-Control Studies (with α = 0.05, 1 − β = 0.80, and Case-Control Ratio 1:2)

	Parameters Estimated from Study	Study Design Based on RR_i	Study Design Based on PAF_i
Marcus *et al.*[2] study			
Ever-smoked	70.0%	70.0%	70.0%
NAT2 status	52.0%	52.0%	52.0%
RR_{10}	1.0	1.0	1.0
RR_{01}	1.3	1.0	1.0
RR_{11}	1.7	2.0	1.3
RR_i	1.3	2.0	1.3
PAF_i	10.9%	26.7%	10%
Sample size		558	3,328
Psaty *et al.*[5] study			
HRT	37.4%	37.4%	37.4%
PT mutation	1.8%	1.8%	1.8%
RR_{10}	0.9	1.0	1.0
RR_{01}	1.5	1.0	1.0
RR_{11}	10.9	10.0	17.5
RR_i	8.1	10.0	17.5
PAF_i	6.2%	5.7%	10%
Sample size		312	213

Examples

We use two case-control studies of gene-environment interaction to illustrate the use of parameter sets based on RR_i and PAF_i in determining sample size. One example represents a common exposure and a common disease-susceptibility genotype with a weak interaction effect. The other example represents a common exposure and a rare disease-susceptibility genotype with a strong interaction effect (Table 1).

Marcus *et al.*[2] conducted a meta-analysis of cigarette smoking, N-acetyltransferase 2 acetylation status (NAT2) and risk for bladder cancer. The study reviewed 16 datasets, including some that lack control subjects. We selected six datasets from European countries with complete case and control subjects (three from England, two from Germany, and one from Denmark) to estimate the parameters needed to calculate sample size. From these data, the estimated prevalence of ever having smoked was 70% (P_e), the prevalence of the NAT2 slow acetylation genotype was 52% (P_g), RR_{10} = 1.0 (CI = 0.7–1.4), RR_{01} = 1.3 (0.9–1.9), and RR_{11} = 1.7 (1.3–2.4). The estimate of RR_i from these data is 1.3, and the associated PAF_i is 10.9% (Table 1, Marcus *et al.*[2] study).

An investigator who wishes to do a similar study might use RR_i to estimate sample size and assume a type I interaction with P_e = 0.7 and P_g = 0.52. Under these conditions, 558 cases would be necessary to detect RR_i = 2 (α = 0.05, 1−β = 0.80, case-control ratio = 2). This is a reasonable number of cases to enroll for a common disease, but the association produces a PAF_i of 26.7%, which the investigator might consider implausibly large for a single interaction effect in a common disease. The investigator might, therefore, re-estimate sample size by

assuming that $PAF_i = 10\%$ is reasonable *a priori*. On this basis, a sample size of 3,328 would be required, and the power would be sufficient to detect an RR_i as small as 1.3. Because the investigator's *a priori* assumptions of $RR_i = 2.0$ and $PAF_i = 10\%$ correspond to quite distinct states of nature, the investigator would need to reexamine the basis for those assumptions to calculate the sample size.

Psaty *et al*. studied hormone replacement therapy, prothrombotic mutation (20210G→A), and the risk for myocardial infarction in postmenopausal women.[5] Among women with hypertension, the estimated prevalence of hormone replacement therapy was 37.4%, the frequency of the prothrombotic mutation (20210G→A) was 1.8%, $RR_{10} = 0.9$ (CI = 0.6–1.4), $RR_{01} = 1.5$ (0.3–7.7), and $RR_{11} = 10.9$ (2.2–55.2). The estimate from these data for RR_i is 8.1 and for PAF_i is 6.2% (Table 1, Psaty *et al*.[5] study).

Suppose that the primary concern of an investigator who wishes to do a similar study is the public health importance of the association. The investigator wants to look for a type I gene-environment interaction with $PAF_i = 10\%$ or more, which she believes is reasonable *a priori*. The number of cases required is 215, but the corresponding $RR_i = 17.5$. This RR_i value may be unrealistically high for the interaction concerned, and the investigator would be well advised to reevaluate the state of nature assumed for the study design. If $RR_i = 10$ were more reasonable *a priori*, the number of cases required would be greater (312) and the associated PAF_i smaller (5.7%). The PAF_i is only moderate in this instance despite the strong interaction effect because the frequency of the susceptibility genotype is low ($P_g = 1.8\%$).

Discussion

Attributable risk estimates provide a public health dimension to the appraisal of risks and an important link between disease causality and public health action.[25] Two recent editorials have, therefore, advocated more frequent use of PAF in epidemiologic studies.[25,26] We have extended the concept of population attributable fraction to studies of gene-environment interactions and have shown that PAF is useful in this context as well.

Our findings have implications for designing investigations of gene-environment interactions. For studies of exposures and susceptible genotypes that are common in a population (for example, P_e and $P_g \geq 30\%$), the associated PAF_i tends to be large even if the strength of the interaction is relatively small (*eg*, $RR_i = 2$ and $PAF_i > 20\%$). From a public health point of view, these studies should receive high priority. In other circumstances, when both the exposure and the susceptible genotype are infrequent in the population, designing a study to identify a substantial attributable risk (*eg*, $PAF_i > 10\%$) might require an interaction effect (RR_i) that is too strong to be biologically plausible. Estimating sample size based on a less extreme RR_i and a lower PAF_i would lead to a more realistic study design but would require more subjects. Even with a reasonably strong interaction effect ($RR_i = 5$), the PAF_i is small (<1%) if the exposure and susceptible genotype are both uncommon (P_e and $P_g <$ 5%). In general, for any interaction of reasonable strength as measured by RR_i, the PAF_i tends to be small if either the prevalence of exposure or the frequency of the susceptible genotype is rare. Even for a strong interaction effect (such as the example of hormone replacement therapy, prothrombotic mutation and the risk for nonfatal myocardial infarction), the PAF_i is relatively small because the susceptible genotype is uncommon in the population. As the prevalence of exposure and the susceptible genotype frequency increase to intermediate values, the PAF_i increases, but a larger sample size is needed to detect the interaction.

Consideration of both RR_i and PAF_i in study design provides investigators with additional insight in making an informed choice about the feasibility, biological plausibility and public health relevance of a study. The fixed mathematic relation between RR_i and PAF_i gives investigators a way to reconcile their intuitive assessment of a measure of effect based on relative odds ratio (RR_i) with one based on public health impact (PAF_i).

When estimation of sample size is based on RR_i as a measure of the strength of interaction, the estimates of PAF_i assume that no confounding exists between exposure, genotype and disease, and the same is true when PAF_i is used as the basis for the calculations. Studies have proposed various formulas to calculate PAF, some of which take into account the effects of confounding.[18] In the absence of confounding, these calculations are equivalent.

Specification of the state of nature to use in estimating the sample size for a study of gene-environment interactions is complex. The choice should be realistic, practical and biologically plausible, and it should also embody public health importance and scientific interest. We considered only three types of gene-environment interactions in which the $PAF_i > 0$. However, the value of PAF_i will be negative if $RR_{11} < RR_{10}\, RR_{01}$ (and $RR_i < 1$). The value of PAF_i under these circumstances can approach negative infinity, but the meaning of such negative values of PAF_i is unknown.

There is a substantial difference between the interpretation of a positive PAF_i value and the interpretation of a conventional attributable fraction calculated for a single exposure variable. PAF_i cannot be interpreted as the proportion of disease cases in the population that would be prevented if both the exposure and susceptible genotype were eliminated. Eliminating the environmen-

tal exposure alone would completely eliminate the effect of the interaction as well as the effect of the environmental exposure on people with other genotypes. In principle, eliminating the susceptible genotype without altering the environmental exposure would also eliminate the interactive effect, but it is not appropriate to consider eliminating a susceptible genotype because this implies elimination of the people who carry that genotype, or at least preventing them from reproducing. The focus must be on elimination or prevention of the environmental exposures. Greenland and Robins have provided additional insights into and cautions about interpretation of PAF.[10]

The number of cases that can be prevented by eliminating an exposure varies among types of gene-environment interactions.[24] For example, for type I interactions, where neither the susceptible genotype alone nor the exposure alone causes excess risk ($RR_{10} = RR_{01} = 1$) but their joint occurrence does ($RR_{11} > 1$), elimination of the environmental exposure would prevent all cases caused by either the genetic susceptibility or the environmental risk factor. For type II interactions, where $RR_{01} = 1$, $RR_{10} > 1$ and $RR_{11} > RR_{10}$, elimination of environmental exposure would prevent all cases resulting from the environmental exposure, regardless of genotype ($PAF_e + PAF_i$). Similar interpretations would apply to other types of gene-environment interactions. In addition, if a given environmental risk factor interacts with susceptibility genes for more than one disease, *eg*, cigarette smoking, NAT2 and bladder cancer[2] or cigarette smoking, CYP1A1 polymorphisms and breast cancer,[26] elimination of the environmental risk factor (in these examples, smoking) would prevent all cases of every disease that results from interactions with that environmental exposure. This could greatly amplify the public health impact of eliminating the environmental exposure.

In general, if $PAF_i > 0$, then more cases of the disease could be prevented by eliminating the exposure in 100 people with the susceptible genotype than by eliminating the same exposure in 100 people in the population as a whole. In other words, the proportion of the disease that is attributable to the gene-environment interaction (PAF_i) provides an estimate of the public health bonus that could be achieved by eliminating the exposure among those with the susceptible genotype.

References

1. Yang Q, Khoury MJ. Evolving methods in genetic epidemiology. III. Gene-environment interaction in epidemiologic research. *Epidemiol Rev* 1997;19:33–43.
2. Marcus PM, Hayes RB, Vineis P, *et al*. Cigarette smoking, N-acetyltransferase 2 acetylation status, and bladder cancer risk: a case-series meta-analysis of a gene-environment interaction. *Cancer Epidemiol Biomarkers Prev* 2000;9:461–467.
3. Feng D, Tofler GH, Larson MG, *et al*. Factor VII gene polymorphism, factor VII levels, and prevalent cardiovascular disease: the Framingham Heart Study. *Arterioscler Thromb Vasc Biol* 2000;20:593–600.
4. Bellamy R. Evidence of gene-environment interaction in development of tuberculosis. *Lancet* 2000;355:588–589.
5. Psaty BM, Smith NL, Lemaitre RN, *et al*. Hormone replacement therapy, prothrombotic mutations, and the risk of incident nonfatal myocardial infarction in postmenopausal women. *JAMA* 2001;285:906–913.
6. Beaty TH, Maestri NE, Hetmanski JB, *et al*. Testing for interaction between maternal smoking and TGFA genotype among oral cleft cases born in Maryland 1992–1996. *Cleft Palate Craniofac J* 1997;34:447–454.
7. Hobbs CA, Sherman SL, Yi P, *et al*. Polymorphisms in genes involved in folate metabolism as maternal risk factors for Down syndrome. *Am J Hum Genet* 2000;67:623–630.
8. Rothman KJ, Greenland S, Walker AM. Concepts of interaction. *Am J Epidemiol* 1980;112:467–470.
9. Kleinbaum DG, Morgenstern H, Kupper LL. *Epidemiologic Research: Principles and Quantitative Methods*. Belmont, CA: Lifetime Learning Publications, 1982.
10. Greenland S, Robins JM. Conceptual problems in the definition and interpretation of attributable fractions. *Am J Epidemiol* 1988; 128:1185–1197.
11. Rothman KJ, Greenland S. *Modern Epidemiology*. 2nd ed. Philadelphia: Lippincott-Raven, 1998.
12. Rothman KJ. *Modern Epidemiology*. 1st ed. Boston: Little Brown and Company, 1986.
13. Hwang SJ, Beaty TH, Liang KY, Coresh J, Khoury MJ. Minimum sample size estimation to detect gene-environment interaction in case-control designs. *Am J Epidemiol* 1994;140:1029–1037.
14. Foppa I, Spiegelman D. Power and sample size calculations for case-control studies of gene-environment interactions with a polytomous exposure variable. *Am J Epidemiol* 1997;146:596–604.
15. Garcia-Closas M, Lubin JH. Power and sample size calculations in case-control studies of gene- environment interactions: comments on different approaches. *Am J Epidemiol* 1999;149:689–692.
16. Gauderman WJ. Sample size requirements for matched case-control studies of gene-environment interaction. *Stat Med* 2002;21:35–50.
17. Browner WS, Newman TB. Sample size and power based on the population attributable fraction. *Am J Public Health* 1989;79:1289–1294.
18. Rockhill B, Newman B, Weinberg C. Use and misuse of population attributable fractions. *Am J Public Health* 1998;88:15–19.
19. Levin M. The occurrence of lung cancer in man. *Acta Union International Contra Cancrum* 1953;9:531–541.
20. Pearce N. Analytical implications of epidemiological concepts of interaction. *Int J Epidemiol* 1989;18:976–980.
21. Botto LD, Khoury MJ. Commentary: facing the challenge of gene-environment interaction: the two-by-four table and beyond. *Am J Epidemiol* 2001;153:1016–1020.
22. Yang Quanhe, Khoury MJ, Flanders WD. Sample size requirements in case-only designs to detect gene-environment interaction. *Am J Epidemiol* 1997;146:713–720.
23. Smith PG, Day NE. The design of case-control studies: the influence of confounding and interaction effects. *Int J Epidemiol* 1984; 13:356–365.
24. Khoury MJ, Beaty TH, Cohen BH. *Fundamentals of Genetic Epidemiology*. *Monographs in Epidemiology and Biostatistics*. Version 22. New York: Oxford University Press, 1993.
25. Northridge ME. Public health methods–attributable risk as a link between causality and public health action. *Am J Public Health* 1995;85:1202–1204.
26. Walter SD. Attributable risk in practice. *Am J Epidemiol* 1998; 148:411–413.
27. Bartsch H, Nair U, Risch A, Rojas M, Wikman H, Alexandrov K. Genetic polymorphism of CYP genes, alone or in combination, as a risk modifier of tobacco-related cancers. *Cancer Epidemiol Biomarkers Prev* 2000;9:3–28.

Appendix

Calculation of Sample Size Required to Detect a Gene-Environment Interaction Producing a Given PAF_i in a Case-Control Study

As in any estimate of sample size required for a study, the investigator must begin by specifying the state of nature for the proposed hypothesis and its alternative. If the desired effect size is to be specified in terms of PAF_i, a set of parameters such as $\{P_{00}, P_{01}, P_{11}, RR_{10}, RR_{01}, PAF_i\}$ that includes PAF_i must be used. (Definitions of the notation used here are provided in the Methods section of the text.) PAF_i and RR_i can be interconverted using Eqs 3 and 5 from the Methods section, so expressing the state of nature in terms of PAF_i can be accomplished by arithmetic manipulation of any standard parameterization of the interactive effect.

The state of nature must then be translated into cell probabilities in a $2 \times 2 \times 2$ table for the case-control study under the null hypothesis (no gene-environment interaction) and its alternative. The expected probability distributions are shown in the Appendix table. The cell probabilities for this table can be defined as follows:

$$\pi_{111} = (P_{11}RR_{10}RR_{01}RR_{11})/T$$
$$\pi_{110} = (P_{10}RR_{10})/T$$
$$\pi_{101} = (P_{01}RR_{01})/T$$
$$\pi_{100} = P_{00}/T$$
$$\pi_{011} = P_{11}$$
$$\pi_{010} = P_{10}$$
$$\pi_{001} = P_{01}$$
$$\pi_{000} = P_{00}$$

where

$$T = P_{00} + P_{01}RR_{01} + P_{10}RR_{10} + P_{11}RR_{10}RR_{01}RR_{11}$$

Suppose the case-control study has n cases and n controls. The variance of the logarithm of RR_i under the null hypothesis, V_N, is approximately

$$V_N = v_N/n$$

where

$$v_N = \sum_{i=0,1}\left(\frac{1}{A_i} + \frac{1}{T_{\cdot 1i} - A_i} + \frac{1}{T_{1\cdot i} - A_i} + \frac{1}{T_{\cdot\cdot i} - T_{\cdot 1i} - T_{1\cdot i} - A_i}\right),$$

TABLE 2. Expected Distribution of Cases and Controls for Gene-Environment Interaction in a Case-Control Study

Exposure	Susceptible Genotype	Cases	Controls	Total
+	+	π_{111}	π_{011}	$T_{\cdot 11}$
−	+	π_{101}	π_{001}	$T_{\cdot 01}$
Total		$T_{1\cdot 1}$	$T_{0\cdot 1}$	$T_{\cdot\cdot 1}$
+	−	π_{110}	π_{010}	$T_{\cdot 10}$
−	−	π_{100}	π_{000}	$T_{\cdot 00}$
Total		$T_{1\cdot 0}$	$T_{0\cdot 0}$	$T_{\cdot\cdot 0}$
Grand total		$T_{1\cdot\cdot}$	$T_{0\cdot\cdot}$	T

the quantity used in Eq 4. The corresponding variance under the alternative hypothesis, V_A, is

$$V_A = v_A/n$$

where

$$v_A = \sum_{i=0,1}\left(\frac{1}{\pi_{i11}} + \frac{1}{\pi_{i10}} + \frac{1}{\pi_{i01}} + \frac{1}{\pi_{i00}}\right),$$

a quantity that is also used in Eq 4.

No closed formula is available to calculate the expected cell probabilities under the null hypothesis of no interaction, but the Mantel-Haenszel approximation (R_{MH}) can be used to approximate V_N, as suggested by Smith and Day,[23] where A_i is the solution of:

$$R_{MH} = \frac{A_i(T_{\cdot\cdot i} - T_{\cdot 1i} - T_{1\cdot i} + A_i)}{(T_{\cdot 1i} - A_i)(T_{1\cdot i} - A_i)} \quad (i = 0,1)$$

and

$$R_{MH} = \frac{\sum_{i=0,1}\frac{\pi_{11i}\pi_{00i}}{T_{\cdot\cdot i}}}{\sum_{i=0,1}\frac{\pi_{01i}\pi_{10i}}{T_{\cdot\cdot i}}}$$

A detailed description of sample size estimation to detect an interaction has previously been published.[22,23]

Eq 4 can be used to estimate sample size based on RR_i by setting the normal deviates $Z_{\alpha/2}$ and Z_β to give a two-sided significance test at level α with power $1-\beta$. To use PAF_i to estimate sample size, one can translate any value of PAF_i to the corresponding RR_i using Eq 5, and then substitute this RR_i into Eq 4.